Application of Machine Learning to the Classification of Spectra-like Data

Yue Sun

Department of Software Engineering
University of Szeged

Szeged, 2023

Supervisor:

Sándor Brockhauser, Péter Hegedűs

Summary of the Ph.D. thesis submitted for the degree of Doctor of Philosophy

of the University of Szeged

University of Szeged

PhD School in Computer Science

Introduction

In scientific research, spectroscopy and diffraction experimental techniques are widely used and produce huge amounts of spectral data. Learning patterns from spectra is critical during these experiments. This provides immediate feedback on the actual status of the experiment (e.g., time-resolved status of the sample), which helps guide the experiment and maximize the scientific output. To this end, it is crucial to be able to employ efficient and automated or semi-automated methods capable of extracting scientifically interesting features in the data.

In this thesis, we examine 1D diffraction spectra data from high-pressure X-ray diffraction (XRD) experiments and in particular focus on the identification of phase transitions in samples investigated by XRD. In these experiments, changes in pressure will cause changes in the unit cell [5], which are reflected in the distribution of spectral peaks. During compression and decompression of the sample, the generated time-resolved spectra can be considered as typical time series data. Detect phase transition implies classifying these time-resolved spectra into three different categories, i.e., before, during, and after the phase transition, which is based on the distribution of spectral peaks. Figure 1 shows X-ray scattering curves corresponding to an example dataset collected from FeO powder sample.

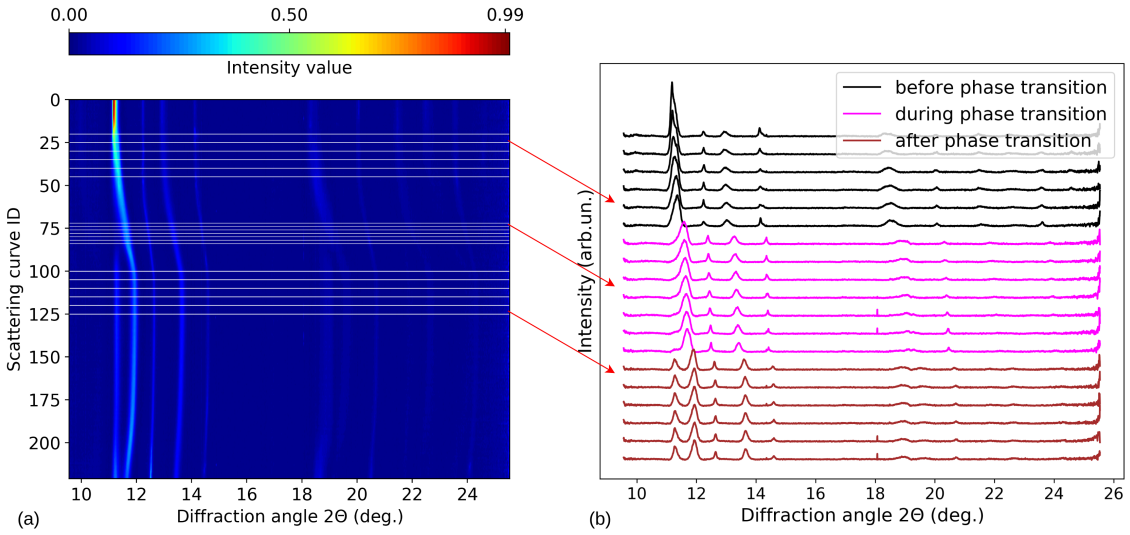

Figure 1: X-ray scattering curves corresponding to an example dataset collected from FeO powder sample. (a) Intensity distribution for different curves collected by applying different pressures drawn as a contour plot. The horizontal lines correspond to representative curves shown in (b). Here, the color black corresponds to the 'before the phase transition', magenta to the transition, and brown to the 'after the phase transition'. In (b) curves are shifted vertically to improve visualization.

Recently, Machine Learning (ML), especially deep learning (DL) models, opens up new avenues for data-driven spectra analysis. They offer great potential for discovering intricate structures and learning representations from high-dimensional data. While promising results have been demonstrated in certain applications of spectra classification within the natural sciences domain, there remains ample room for further exploration and advancement in this field, such as better data representation, generalization and interpretability of models, and efficiency and automation of classification. It is crucial to address these limitations while applying machine learning and deep learning techniques to

1D diffraction spectra classification to ensure reliable and meaningful results. In this thesis, we employ tools (e.g., data analysis and machine learning) from the field of software engineering to tackle the problem of diffraction spectra classification and in particular the identification of phase transitions in samples investigated by x-ray powder diffraction.

Given that spectral classification is the domain-specific problem to be solved, the key features of the spectral data should be considered when designing the ML-based classification models, as shown below.

- The spectra have high dimensionality, with about 4000 features (data points) in each spectral curve. Whereas the number of diffraction files is usually much smaller than the number of features, in our use case there are 60-460 spectral curves in each dataset. This can easily lead to the overfitting of ML models, especially for supervised ML models.

- The most important information is the spectral peaks, e.g. the number, position, shape, and intensity of the peaks reflect the main features of the evaluated sample. These features have local and global dependencies in the diffraction angle dimension and are not independent. Statistical models should focus on peak information while suppressing noise and irrelevant features as much as possible.

- The spectral data were collected during high-pressure experiments. In this case, the distribution of the spectral peaks varies with the pressure (external variable), describing the dynamic process of the phase transition. Thus, spectral data have a strong dependence on external variables, which we uniformly refer to as time dependence.

This dissertation focuses on addressing the challenges outlined above by presenting several different types of ML/DL methods tailored for spectra-like data, with the aim of achieving efficient and automatic classification. The three main results of the thesis are the following:

I. We start with models with manual feature engineering. First, we propose a spectral classification model based on PCA preprocessing and spectral clustering. Then, we design a neural network-based ML model combined with a binned-weighting technique for spectra classification.

II. We introduce several end-to-end supervised classification models that eliminate the need for feature engineering. In particular, the proposed convolutional SCT attention model can learn better representations by modeling global dependencies in spatial, channel, and temporal dimensions. In addition, we introduce a supervised contrast learning framework and interpret the model in terms of traditional spectral descriptors.

III. We propose three self-supervised frameworks to classify 1D spectral data using a minimal amount of labeled data. These frameworks are based on relational reasoning (SpecRR-Net), contrastive learning (SpecMoco-Net), or a combination of the two (SpecRRMoco-Net). In addition, we discuss and validate augmentations relevant to the case study discussed, ensuring the retention of scientifically meaningful information.

I Spectra Classification with Manual Feature Engineering

The contributions of this thesis point are related to applying machine learning methods with feature engineering to spectra classification.

Spectra Classification Model Based on the Principle Component Analysis (PCA) and Spectral Clustering

When considering the application of machine learning methods to spectral data analysis, scientists generally have two tools at their disposal: (i) clustering the data to distinguish between different classes of samples, or (ii) labeling selected data to train a supervised classifier. We start with a spectral clustering method [9], which is typically used in the data exploration phase. Specifically, we present a spectra classification model based on the Principle Component Analysis (PCA) and Spectral Clustering. In this model, PCA is applied to reduce the dimensionality, after which the spectral clustering method is applied. To better explain the model, we investigated the contribution of the original variables in spectra to the principle components in PCA and also tested the relationship between the classification results and the number of PCs (Fig. 2). In addition, in the absence of ground-truth label information, we propose the classification confidence metric for model evaluation. The results in Fig. 2 show show that the method obtains consistently high-precision classification results and that the obtained classification boundaries are very stable, fluctuating only within a small range.

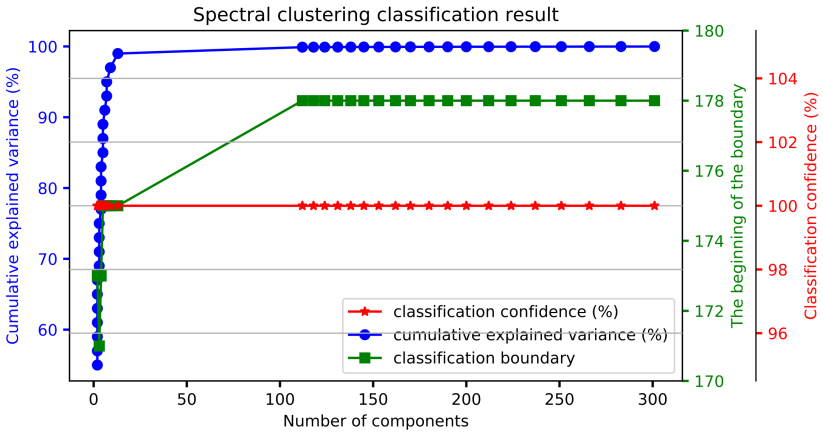

Figure 2: Classification confidence with different explained variance value and PCs.

Despite the high performance we achieved on the example spectral dataset, there are still some limitations in this method. First, determining the appropriate number or density of clusters in the clustering method is a challenge, especially when dealing with complex and high-dimensional spectral data, which often requires fine-tuning certain hyperparameters to obtain accurate results, thus hindering automation performance. Second, the approach relies on PCA for preprocessing, which may lead to the loss of important discriminatory information. Furthermore, this additional step of feature engineering can impede automation, and introduce more complexity into the interpretation of the model.

ML Model Combined With a Binned-Weighting Technique for Spectra Classification

To address these limitations of the clustering method, we further propose a data-driven neural network-based statistical model [7] applied to spectra classification. Neural networks are chosen here due to their ability to learn complex mappings between input and target spaces which makes them perfect for our task. In this model, we transform our problem of finding and distinguishing features in a set of 1D curves where the input is provided by the sequence of the spectral intensity values, into a 2D segmentation problem where we input every point in the spectra with their 2 coordinates (scattering angle, and azimuthally integrated intensity). Since classifying features in overlapping regions is difficult, meaningless, and of low confidence, machine learning models with a binned-weighting technique are proposed to minimize the misclassification of indistinguishable features in overlapping regions. The believability weighting factor is calculated based on the classification accuracy (separability) of the neural network for each bin. The solution proposed here can automatically find regions (or bins) with high separability, as shown in Figure 3. The final classification label for each spectral curve is calculated by the weighted sum of the point-wise classification results assigned by the neural networks in each bin.

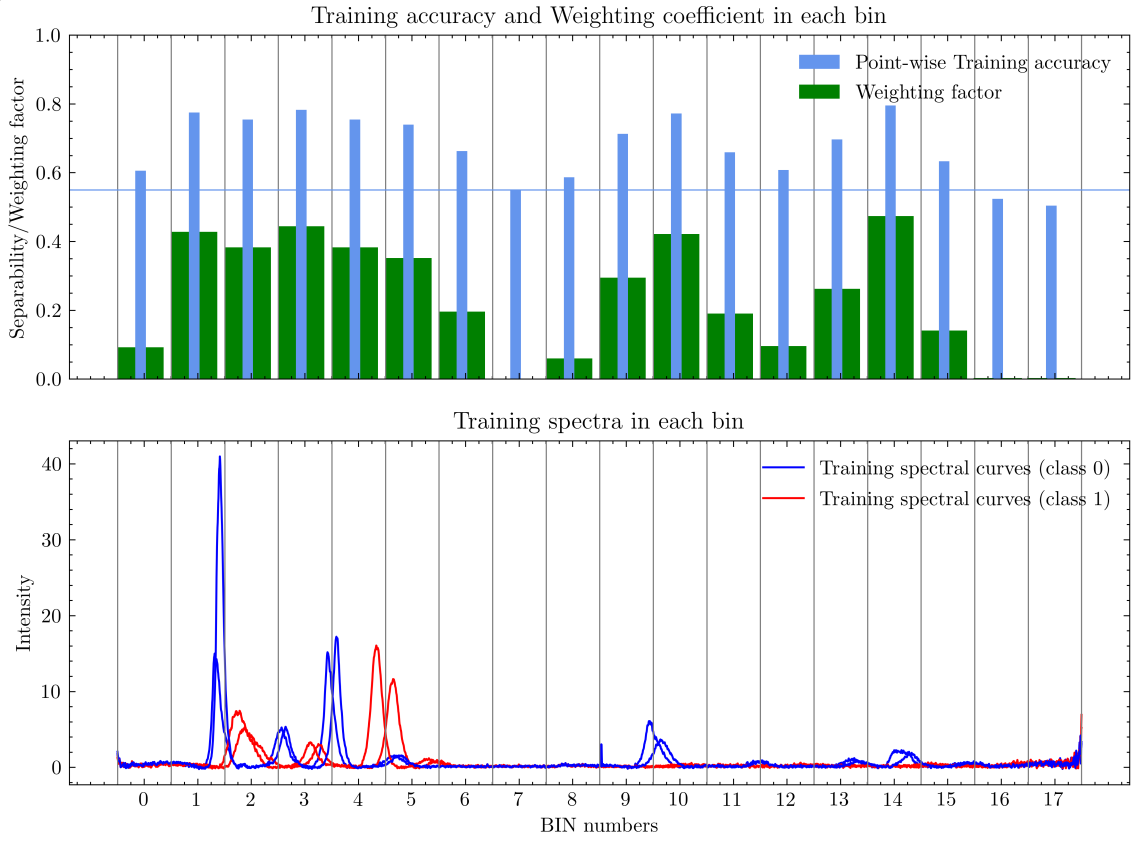

Figure 3: The distribution of point-wise training accuracy (separability), and the believability weighting factor for each bin.

In addition, we also investigated the performance of the model using a different number of bins. The results show that the more bins applied, the higher the classification accuracy, thus the smaller the ambiguous zone. We can conclude that the key to achieving high classification confidence in this approach is to find bins with high separability indices.

The Author's Contributions

The author evaluated and compared clustering methods and chose the Spectral Clustering method to analyze the spectra data. She studied the principal components (PCs) of the PCA method and analyzed the contribution of the original variables to the PC. In addition, she tested the classification confidence with different cumulative explained variance values in the PCA method. Furthermore, the author performed the empirical validation of the neural-network-based ML model combined with a binned-weighting technique for spectra classification and drew some key conclusions. She implemented, analyzed, evaluated, and presented the results. Finally, she provided the analysis script as Jupyter Notebooks for reproducibility.

II End-to-End Deep Learning Methods for Spectra Classification

The contributions of this thesis point are related to the development of end-to-end deep learning models and an open-source interactive software application for spectra classification.

End-to-End Machine Learning Methods for Spectra Classification

To further improve the generality and automation of the classification models, we propose several deep classification models in an end-to-end manner [8], eliminating the need for manual feature engineering. We first extend the neural network model with weighting techniques to an end-to-end binned FCNN (fully connected neural network) with the automatically capturing weighting factors model. With this improvement, the local believability weighting factors of each bin are learned automatically during training and are dynamically responsive to the input data, so the model can automatically prominent features with high separability and suppress indistinguishable features with low separability.

More generally, the spectra classification problem can be regarded as a 1D time series classification problem, and in this setting, the convolutional SCT (spatial-channel-temporal) Attention model (Fig. 4) is proposed. It is a hybrid model of CNN (convolutional neural network) and self-attention architecture, where CNN is used to learn local features while self-attention is calculated across spatial, channel, and temporal dimensions, enabling the model to focus on distinguishable features while suppressing noisy and misleading ones. We also designed and implemented several other classification models tailored for 1D spectra based on other state-of-the-art ML algorithms, such as FCNN, CNN, Resnet, LSTM, and Transformer. These end-to-end solutions bring us closer to achieving efficient and automated classification.

Furthermore, we evaluated and compared the classification performance of these deep end-to-end learning models on an example experimental spectra dataset from multiple perspectives (Fig 5 and 6). The results show that the proposed models in 1D time series classification are superior. In addition, we investigated all the models in the 1D time series classification setting further by a feature importance analysis using gradients backpropagation [6]. In this experiment, we estimate the contribution of each input feature to the classification prediction of each model. The results show that the convolutional SCT attention model can best focus on classification-related features and suppress non-suppressible noise features. In summary, on this basis, this work provides a standard baseline for 1D spectral time

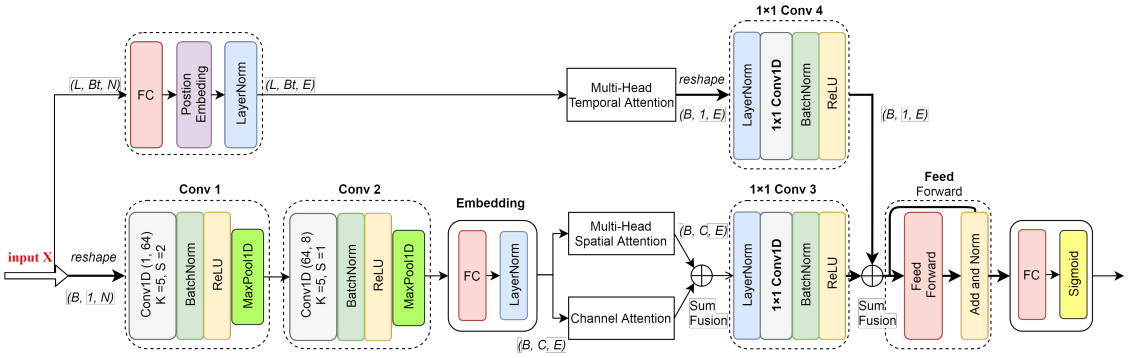

Figure 4: Illustration of convolutional SCT attention network architecture. In this architecture, attention is calculated across spatial, channel, and temporal dimensions.

series classification in an end-to-end manner.

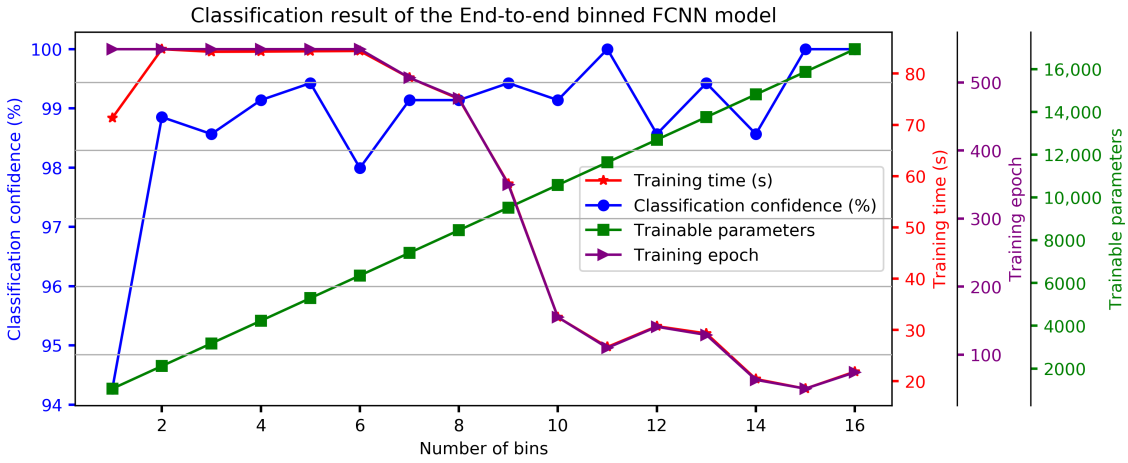

Figure 5: Training information and classification result of the end-to-end binned FCNN with automatically capturing weighting factors model.

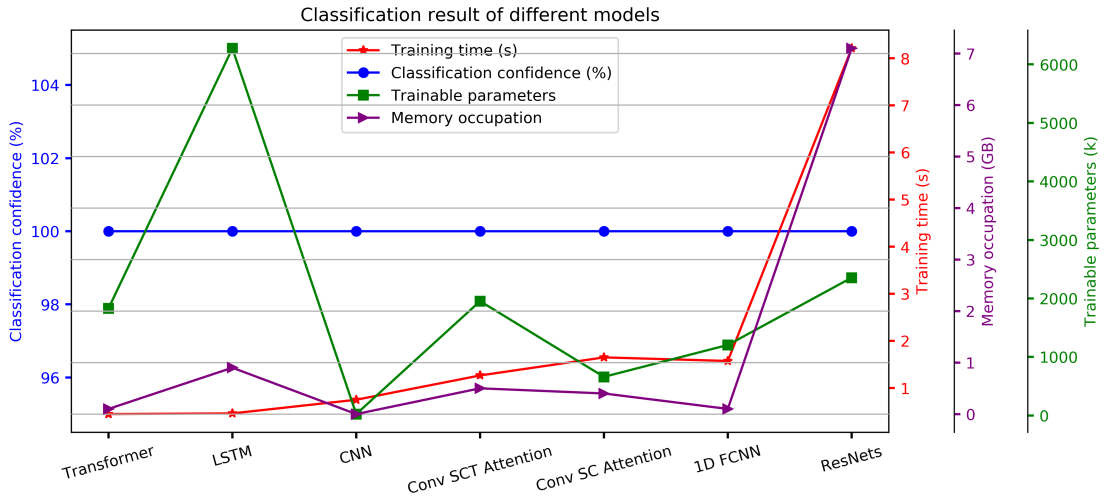

Figure 6: Training information and classification results of these models under the 1D time series data classification problem setting.

Supervised Contrastive Learning for Spectra Classification

We further propose a novel classification framework based on supervised contrastive loss for one-dimensional spectral classification of diffraction spectra, inspired by recent studies on supervised contrastive learning [4]. Instead of optimizing the model solely with cross-entropy loss, we leverage the power of contrastive learning to enable the model to learn more general representations from the data. The backbone encoder utilized in our framework is based on the convolutional SCT attention model described above.

To ensure the effectiveness and robustness of our proposed framework, we conduct experiments on multiple diverse experimental diffraction spectra datasets, rather than relying on a single example dataset for validation. Additionally, we aim to provide insights into the decision-making process of the deep classification model and bridge the gap between complex models and physicists' understanding. Therefore, we interpret the model in terms of traditional spectral mode descriptors that are of particular interest and relevance to physicists, such as peak position, width, and intensity.

Through the analysis and interpretation of the learned representations, we demonstrate that the network has the capability to automatically extract physically meaningful peak information, which is essential for accurate spectral classification tasks. This interpretation allows physicists to gain valuable insights into the model's behavior and align it with their domain expertise, facilitating a better understanding and utilization of the model in practical applications.

An Interactive Machine Learning Application for Spectra Classification

To enhance data analysis, model reuse, and the advancement of new methodologies, the thesis presents an open-source interactive software program [12] specifically designed for spectral classification based on these end-to-end supervised learning methods. By implementing the scripts on the Jupyter Notebook platform and utilizing interactive runtime environments such as Mybinder and Google Colab, the software program facilitates improved reproducibility, as well as effortless evaluation and exploration of data and models.

The Author's Contributions

First, the author conducted an in-depth analysis of diffraction spectra data and identified the need for new classification models tailored specifically for spectra data. Then, the author introduced novel end-to-end supervised classification models with the objective of improving performance and enhancing automation and efficiency by eliminating the need for manual feature engineering. In order to interpret the classification models, the author conducted a feature importance analysis based on gradient backpropagation, providing insights into the significance of different features. In addition, she proposed a spectral classification model based on supervised contrastive learning, which is the first attempt to apply this method to diffraction spectra. Furthermore, the author interpreted the models from multiple perspectives, employing traditional descriptors of diffraction spectral patterns relevant to physicists, such as peak number, peak position, width, and intensity. Lastly, the author's contribution extends to the development of an open-source interactive software program that facilitates these end-to-end approaches to spectral classification.

III Self-Supervised Approaches for Spectra Classification

The contributions of this thesis point are related to the development of more automatic spectral classification models based on self-supervised learning. These models are capable of learning discriminative features and building effective representations, therefore greatly reducing the number of labels required, making a step towards automating the spectral classification process.

Self-Supervised Spectra Classification Models

Although end-to-end supervised classification models, particularly the convolutional SCT attention model, have shown high accuracy in learning data representations and making classification predictions directly from raw data without the need for feature engineering, the data labeling in these methods still hampers the automation process of spectral classification.

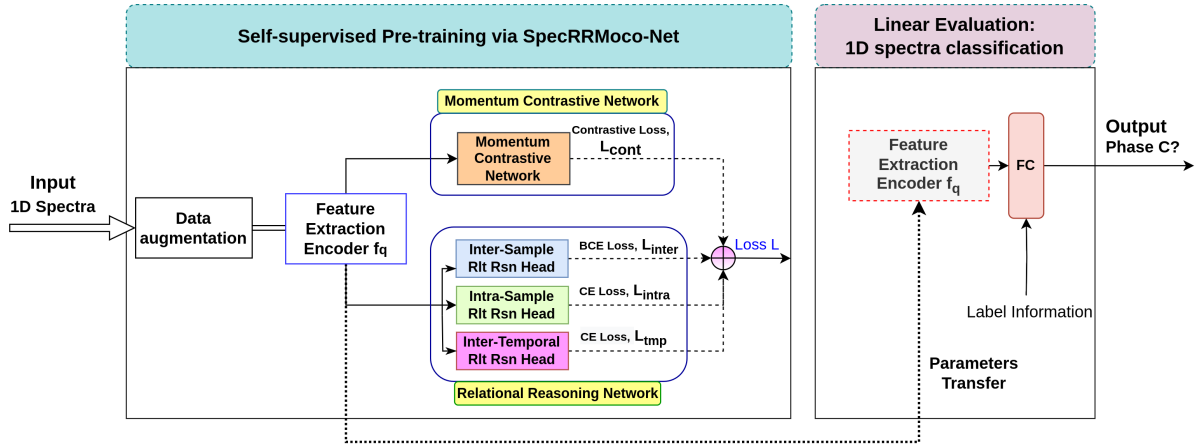

Figure 7: Illustration of the proposed 1D spectra classification framework based on the self-supervised SpecRRMoco-Net, which is a combination of Relational Reasoning Network (SpecRR-Net) and Momentum Contrast Network (SpecMoco-Net). The classification framework consists of two parts, namely pre-training and linear evaluation of downstream spectral classification. In the pre-training stage, the encoder f_q is trained on unlabeled data to build useful representations; in the linear evaluation stage, a small number of labels are used to perform the downstream spectral classification task, where a linear classifier is trained on top of the frozen feature extraction encoder f_q. Specifically, the encoder is training by jointly minimizing the contrastive loss L_{cont} in SpecMoco-Net, inter-sample relational reasoning loss L_{inter}, the intra-sample relational reasoning loss L_{intra}, and the inter-temporal relational reasoning loss L_{tmp} in SpecRR-Net.

To address this limitation, this thesis introduces self-supervised classification frameworks [10] based on self-supervised relational reasoning [2, 13] (SpecRR-Net), self-supervised contrastive learning [1, 3] (SpecMoco-Net), or a combination of both (SpecRRMoco). The pretext tasks and data augmentations proposed in this work are specifically tailored to the scientific problem at hand, preserving physically meaningful information. Additionally, an inter-temporal relational reasoning module is incorporated in SpecRR-Net and SpecRRMoco-Net to capture temporal dependencies in time-resolved spectral data, which significantly improves the quality of learned representations. These self-supervised models have the ability to learn discriminative features and construct effective representations, thus reducing the reliance on a large number of labels and advancing the automation of spectral classification. Moreover,

as a consequence of the reduced number of labels, scientists' time is greatly optimized. Lastly, the thesis emphasizes the importance of selecting appropriate data augmentations that are specifically designed for the particular study case, ensuring the retention of scientifically meaningful information.

Self-Supervised Loss Function and its Validation

The proposed self-supervised loss function in the SpecRRMoco-Net unifies the relational reasoning-based model (SpecRR-Net) and contrastive learning-based model (SpecMoco-Net), and is a linear combination of these two networks. To our best knowledge, this is the first time they have been combined together. We report on an ablation study on the combination coefficient in the loss function, which was performed to understand its impact on learning data representations.

Comparison of These Self-Supervised Methods

We compare these three self-supervised classification models from several perspectives, including performance in downstream classification tasks, as well as clustering power analysis, such as visualization of embedded features learned by each model, and evaluation by Silhouette Score and Mutual Information metrics. We concluded that SpecRRMoco-Net shows superior performance by benefiting from contrastive learning and relational inference learning. Furthermore, by comparing with a state-of-the-art self-supervised relational reasoning model (SpecSelfTime) without an inter-temporal relational inference module, we also emphasize the importance of this module we introduced. Table 1 shows the average classification precision/recall of SpecRRMoco-Net, SpecRR-Net, SpecMoco-Net, and SpecSelfTime, on Fe datasets and FeO datasets.

	Model	**Fe**		**FeO**	
		Precision	**Recall**	**Precision**	**Recall**
2.8% labels	SpecSelfTime	98.6 ± 0.2	98.5 ± 0.2	78.6 ± 3.2	80.0 ± 4.3
	SpecRR-Net	99.2 ± 0.3	99.1 ± 0.3	91.6 ± 6.2	90.7 ± 5.4
	SpecMoco	98.3 ± 1.1	97.9 ± 1.8	93.8 ± 3.7	93.2 ± 4.1
	SpecRRMoco	99.6 ± 0.2	99.6 ± 0.2	96.3 ± 3.2	96.5 ± 2.4
10% labels	SpecSelfTime	98.6 ± 0.5	98.5 ± 0.5	80.5 ± 2.7	82.0 ± 3.9
	SpecRR-Net	99.1 ± 0.2	99.0 ± 0.2	96.9 ± 0.8	95.8 ± 1.1
	SpecMoco	99.3 ± 0.3	99.3 ± 0.3	94.1 ± 0.7	93.7 ± 0.7
	SpecRRMoco	99.6 ± 0.2	99.5 ± 0.2	97.1 ± 1.8	96.9 ± 1.3

Table 1: Classification results measured in terms of weighted precision and recall using different self-supervised methods. For each method, the classification results are reported with amounts of labels corresponding to either 2.8% or 10% of the total collected data.

Ablation Studies on the Data Augmentation

We also conducted an ablation study on data augmentations performed in order to evaluate their impact on the models' performances, emphasizing the need to tailor data augmentation approaches to the specific case study. Figure 8 shows the linear evaluation under different data augmentation techniques individually or in combination. In particular, the best result is achieved when 'magnitude warping', is combined with 'diffraction angle warping'. It is worth noting that the data augmentation

involved in the self-supervised learning approaches respects the fundamental invariance of the physical problem. We also observed that the sequence in which data augmentation techniques are applied may have an impact on the results, as different orders yield distinct enhanced data due to the inherent randomness associated with each data enhancement technique. As data augmentation is domain-specific, it must be customized for data sets from different research areas.

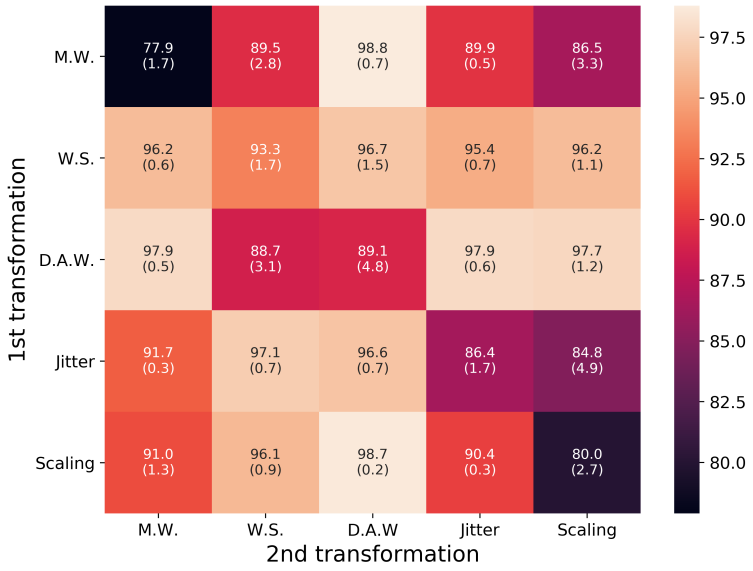

Figure 8: Ablation study on data augmentation techniques. Results for magnitude warping (M.W.), window slicing (W.S.), diffraction angle warping (D.A.W.), jittering (Jitter), and scaling data augmentation techniques are reported. The figure shows the average classification accuracy and standard deviation (values in parentheses) for 20 runs with 2.8% of labeled data. In addition to this, diagonal elements indicate the use of only one data augmentation technique, while other non-diagonal entries indicate the combination of two data augmentation techniques. The color scale represents the classification accuracy.

An Open-Source Interactive Software Program Based on Self-Supervised Learning Classification Models

Similar to the open-source interactive software program based on end-to-end supervised deep classification models, we provide an interactive spectral classification program based on SpecRRMoco-Net.

It is worth noting that since self-supervised learning includes pre-training and linear evaluation, its training process and preparation of training/test data are different from supervised learning models. For example, pre-training in self-supervised learning methods is based on unlabeled data, and their generation of training/test data and definition of excuse tasks involve data augmentation techniques.

The Author's Contributions

The author's contribution to this thesis point encompasses several key aspects. First, the author proposed novel self-supervised classification models, including the design of pretext tasks and the selection of data augmentation, to achieve improved automation and efficiency in spectral data classification. Second, she conducted an ablation study on the selection of data augmentation techniques, crucial

to ensure that scientifically meaningful information is retained. Furthermore, she implemented and extensively validated these models using several experimental spectral datasets for a comprehensive evaluation from multiple perspectives. In addition, the analysis and discussion of the results is also a contribution of hers. Finally, she provided an interactive software program for spectral classification, based on these self-supervised methods.

Summary

The main results presented in the thesis are related to the employment of techniques such as data analysis and machine learning in the field of software engineering to develop efficient, automatic, and interpretable ML/DL-based classification models for 1D spectral classification. The models proposed in this thesis take careful consideration of the unique features of spectral data.

The contributions of the thesis are grouped into three major thesis points. We start with models with manual feature engineering. First, we proposed a spectral classification model based on PCA and spectral clustering, which are typically used in the data exploration phase. Then, we designed a neural network-based ML model combined with a binned-weighting technique for spectra classification to minimize misclassifications of indistinguishable features in overlapping regions of different spectra. In this model, the spectra classification problem is transformed into a 2D segmentation problem. However, the high classification performance of both methods involves feature engineering, which poses a challenge to the automation of the models and may limit their generalizability when applied to different datasets.

Next, to achieve better classification performance and to get rid of manual feature engineering, we proposed several deep classification models in an end-to-end manner. We focus on two proposed classification models, namely the end-to-end binned FCNN with automatically capturing weighting factors model when viewed as a 2D space segmentation problem and the convolutional SCT attention model when viewed as a 1D time series classification problem. And several other end-to-end model structures based on FCNN, CNN, ResNets, Transformer, and LSTM are explored. Finally, we evaluated and compared the performance of these classification models from multiple perspectives. To further increase the generability of the classification model, we introduce a supervised contrastive learning framework for 1D spectral representation and classification. Additionally, we provided an interpretation of this classification model in terms of traditional descriptors of spectral data.

Finally, we proposed three self-supervised frameworks to classify 1D spectral data using a minimal amount of labeled data. These frameworks are based on relational reasoning (SpecRR-Net), contrastive learning (SpecMoco-Net), or a combination of the two (SpecRRMoco-Net). They are capable of learning discriminative features and building effective representations, therefore greatly reducing the number of labels required, making a step towards automating the spectral classification process. We demonstrate the importance of a proper choice of data augmentations, which must be tailored for the specific case of study to ensure the retention of scientifically meaningful information. Finally, we provide a convenient software platform for spectral classification based on end-to-end supervised classification models and self-supervised classification models.

Here, we also summarize the publications related to the various thesis points in Table 2.

$\mathcal{N}o.$	[9]	[7]	[8]	[11]	[12]
I/1.	•	•			
II/2.			•		•
III/3.				•	•

Table 2: Thesis contributions and supporting publications

Acknowledgements

I would like to thank my supervisor Sándor Brockhauser, Péter Hegedűs for guiding me in my studies and research, without whom I would not have been able to conduct my research. I am very grateful to them for respecting my ideas and giving me the freedom to explore and think. I joined the European XFEL Data Analysis Group in October 2021 and I would also like to express my sincere thanks to Dr. Luca Gelisio and Dr. Danilo Enoque Ferreira de Lima, whom I regard as my second mentor, for their excellent supervision and support during this period. I would like to thank Christian Plueckthun and Zuzana Konopkova at European XFEL for providing the HED experimental spectral data and valuable discussions and interpretations. I would also like to thank Dr. Liu Shan from DESY (Hamburg, Germany) for being a mentor and sharing her knowledge and skills with me to help me plan my future academic and life path. I am also very grateful to Dr. Jun Zhu and Dr. Ye Chen for supporting and helping me in my academic career. My many thanks also go to my friends, namely Dr. Zhuoni Qian, Dr. Juncheng E, Dr. Jiawei Yan, Dapeng Li, Dr. Zihan Zhu, Dr. Junjie Guo, Dr. Weilun Qin, Dr. Tianyun Long, Dr. Xinchao Huang, Dr. Hao Wang, Dr. Bihan Wang, and Dr. Minxue Tang for their sincere support. Finally, I would like to thank my boyfriend, Xiayun Pan, and my family for their unwavering support and encouragement throughout my PhD. Their constant support and trust in me has been a source of strength and motivation. I am deeply grateful for their support and love. It should also be mentioned that this research was supported by China Scholarship Council (CSC, No. 201904890020), and the European XFEL.

Yue Sun, July 2023

References

[1] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297, *2020.*

[2] Haoyi Fan, Fengbin Zhang, and Yue Gao. Self-supervised time series representation learning by inter-intra relational reasoning. arXiv preprint arXiv:2011.13548, *2020.*

[3] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, *pages 9729–9738, 2020.*

[4] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. Advances in neural information processing systems, *33:18661–18673, 2020.*

[5] Christian Plückthun. Investigating the effect of the compression rate on the kinetic response of diamond anvil cell experiments. Technical report, Universität Rostock, 2022.

[6] Marc Rußwurm and Marco Körner. Self-attention for raw optical satellite time series classification. ISPRS journal of photogrammetry and remote sensing, *169:421–435, 2020.*

[7] Yue Sun and Sandor Brockhauser. Machine learning applied for spectra classification in x-ray free electorn laser sciences. Data Science Journal, *21(1), 2022.*

[8] *Yue Sun, Sandor Brockhauser, and Péter Hegedűs. Comparing end-to-end machine learning methods for spectra classification.* Applied Sciences, *11(23):11520, 2021.*

[9] *Yue Sun, Sandor Brockhauser, and Péter Hegedűs. Machine learning applied for spectra classification. In* Computational Science and Its Applications–ICCSA 2021: 21st International Conference, Cagliari, Italy, September 13–16, 2021, Proceedings, Part IX 21, *pages 54–68. Springer, 2021.*

[10] *Yue Sun, Sandor Brockhauser, Péter Hegedűs, Christian Plückthun, Luca Gelisio, and Danilo Enoque Ferreira de Lima. Application of self-supervised approaches to the classification of x-ray diffraction spectra during phase transitions.* Scientific Reports, *13(1):9370, 2023.*

[11] *Yue Sun, Sandor Brockhauser, Péter Hegedűs, Christian Plückthun, Luca Gelisio, and Danilo Enoque Ferreira de Lima. Application of self-supervised approaches to the classification of x-ray diffraction spectra during phase transitions.* Scientific Reports, *13(1):9370, 2023.*

[12] *Yue Sun, Christian Plückthun, Sandor Brockhauser, and Péter Hegedűs. An interactive machine learning application for spectra classification. In* Computational Science and Its Applications–ICCSA 2023: 23st International Conference, *pages Accepted, to appear. IEEE Computer Society, 2023.*

[13] *Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In* Proceedings of the IEEE conference on computer vision and pattern recognition, *pages 1199–1208, 2018.*

Declaration

In the PhD dissertation of Yue Sun entitled 'Application of Machine Learning to the Classification of Spectra-like Data', Yue Sun's contribution was decisive in the following results:

1. Spectra classification model based on the Principle Component Analysis (PCA) and Spectral Clustering

The author proposed a spectral classification model based on PCA and spectral clustering. She studied the principal components (PCs) of the PCA method and analyzed the contribution of the original variables to the PC. She tested the classification confidence with different cumulative explained variance values in the PCA method. She implemented, analyzed, evaluated, and presented the results. Finally, she provided the analysis script as Jupyter Notebooks for reproducibility.

This result is related to Thesis point 1 (Chapter 3) and publication 4.

2. ML model combined with a binned-weighting technique for spectra classification.

The author designed a neural-network-based ML model combined with a binned-weighting technique for spectra classification to minimize misclassifications of indistinguishable features in overlapping regions of different spectra. She performed the empirical validation of the model and drew some key conclusions. She implemented, analyzed, evaluated, and presented the results. The analysis script based onJupyter Notebooks has been open-sourced for reproducibility. This work was also been selectedas a use case in project 'The Photon and Neutron Open Science Cloud (PaNOSC)'.

This result isrelatedto Thesis point 1 (Chapter 4) and publication 3.

3. End-to-End Machine Learning Methods for Spectra Classification

The author conducted an in-depth analysis of diffraction spectra data and identified the need for new classification models tailored specifically for spectra data. Then, the author introduced novel end-to-end supervised classification models with the objective of improving performance and enhancing automation and efficiency by eliminating the need for manual feature engineering. She first extended the previous ML model combined with a binned-weighting technique to the end-to-end binned Fully Connected Neural Network (FCNN) with the automatically capturing weighting factors model. Under the setting of 1D time series classification, she proposedthe convolutional SCT attention model, which can learn better representations by modeling the global dependence of spatial, channel, and temporal dimensions to effectively suppress noisy features with low separation. In addition to this, shedesigned and implemented several different classification models based on state-of-the-art deep learningmodels for spectral classification. She evaluated and compared the performance of these classification models from multiple perspectives. In order to betterinterpret the classification models, sheconducted a feature importance analysis based on gradient backpropagation, providing insights into the significance of different features.

This result isrelatedto Thesis point 2 (Chapter 5) and publication 2 and 5.

4. Supervised contrastive learning for spectra classification

The author introduced a supervised contrastive learning framework with data augmentation for 1D spectral representation and classification, a first attempt in diffraction spectroscopy applications. Furthermore, the author interpreted the models from multiple perspectives, employing traditional descriptors of diffraction spectral patterns relevant to physicists, such as peak number, peak position, width, and intensity. In addition, she implemented, analyzed, evaluated, and presented the results.

This result is related to Thesis point 2 (Chapter 6).

5. Self-Supervised Approaches for Spectra Classification

The author proposed novel self-supervised classification models, including the design of pretext tasks and the selection of data augmentation, to achieve improved automation and efficiency in spectral data classification. The self-supervised classification frameworks are based on self-supervised relational reasoning (SpecRR-Net), self-supervised contrastive learning (SpecMoco-Net), or a combination of both (SpecRRMoco). These self-supervised models have the ability to learn discriminative features and construct effective representations, thereby reducing the reliance on the number of labels and advancing the automation of spectral classification. In particular, the proposed SpecRRMoco-Net unifies the model based on relational reasoning (SpecRR-Net) and the model based on contrast learning (SpecMoco-Net). To the best of our knowledge, this is the first time that these two self-supervised learning methods are combined. In addition, in SpecRR-Net and SpecRRMoco-Net, she proposed an inter-temporal relational reasoning module to capture temporal dependencies in time-resolved spectral data, which significantly improves the quality of learned representations.

She also conducted an ablation study on the selection of data augmentation techniques, crucial to ensure that scientifically meaningful information is retained. Furthermore, she implemented and extensively evaluated these models from multiple perspectives based on several experimental spectral data sets. In addition, the analysis and discussion of the results is also a contribution of hers.

This result is related to Thesis point 3 (Chapter 7) and publication 1 and 5.

6. An Interactive Machine Learning Application for Spectra Classification.

To enhance data analysis, model reuse, and the advancement of new methodologies, the author implemented and presented an open-source interactive software program specifically designed for spectral classification. This program incorporates two types of models introduced in the thesis: end-to-end supervised learning methods (Chapter 5) and self-supervised classification methods (Chapter 7). By implementing the scripts on the Jupyter Notebook platform and utilizing interactive runtime environments such as Mybinder and Google Colab, the software program facilitates improved reproducibility, and allows for efficient evaluation and exploration of data and models.

This result relates to Thesis point 2 and 3 and publication 5.

The main thesis points are listed below:

Thesis point 1: Spectra classification with manual feature engineering.

Thesis point 2: End-to-End Deep Learning Methods for Spectra Classification.

Thesis point 3: Self-Supervised Approaches for Spectra Classification.

The publications are listed below:

1. Sun, Y., Brockhauser, S., Hegedűs, P., Plückthun, C., Gelisio, L. and Ferreira de Lima, D.E., 2023. Application of self-supervised approaches to the classification of X-ray diffraction spectra during phase transitions. *Scientific Reports*, *13*(1), p.9370.
2. Sun, Y., Brockhauser, S. and Hegedűs, P., 2021. Comparing End-to-End Machine Learning Methods for Spectra Classification. *Applied Sciences*, *11*(23), p.11520.
3. Sun, Y. and Brockhauser, S., 2022. Machine Learning Applied for Spectra Classification in X-ray Free Electorn Laser Sciences. *Data Science Journal*, *21*(1).
4. Sun, Y., Brockhauser, S. and Hegedűs, P., 2021. Machine Learning Applied for Spectra Classification. In *Computational Science and Its Applications–ICCSA 2021: 21st International Conference, Cagliari, Italy, September 13–16, 2021, Proceedings, Part IX 21* (pp. 54-68). Springer International Publishing.
5. Sun, Y., Brockhauser, S. and Hegedűs, P., 2023. An Interactive Machine Learning Application for Spectra Classification. In *Computational Science and Its Applications–ICCSA 2023: 23st International Conference, 2023, Proceedings,* (Accepted, to appear). Springer International Publishing.

These results cannot be used to obtain an academic research degree, other than the submitted PhD thesis of Yue Sun.

Yue Sun

June 26, 2023

date, signature of candidate, signature of supervisor

The head of the Doctoral School of Computer Science declares that the declaration above was sent to all of the coauthors and none of them raised any objections against it.

July 10, 2023

date, signature of head of Doctoral School

SZEGEDI TUDOMÁNYEGYETEM
INFORMATIKA DOKTORI ISKOLA
SZEGED